3.3 Recovering Population Effects via Regression

In the previous section we saw that, under parametric model assumptions (with randomization), the coefficient from a logistic regression model recovers a conditional odds ratio. Here we consider the question of how we might use this fit to estimate the marginal average treatment effect.

First a few basics. When we fit a logistic (or any other) regression model we are estimating the conditional expectation function

$$\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a).$$

For example in logistic regression we estimate the function μ with

$$\widehat{\mu}_a(x) = \text{expit}(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\beta}_2^{\mathrm{T}} x).$$

In R this function can be evaluated at the observed $(X_1, A_1), ..., (X_n, A_n)$ values with the `predict` command, as in:

```
lrmod <- glm(y ~ x+a, family=binomial)
muhat <- predict(lrmod, type="response")
```

Now recall that under the randomization assumption $A \perp\!\!\!\perp (Y^a, X)$ (together with consistency) the average treatment effect is given by

$$\psi = \mathbb{E}(Y^1 - Y^0) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)\}$$

which suggests the estimator

$$\begin{aligned}\widehat{\psi} &= \mathbb{P}_n\Big\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\Big\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big\}\end{aligned}$$

This estimator is sometimes called the *plug-in*, *g-computation*, or *standardization* estimator. Conceptually, it is taking the estimated conditional effect $\widehat{\mu}_1(x) - \widehat{\mu}_0(x)$ and standardizing it to the (empirical) population distribution of covariates. You can also think of it as "imputing" an estimate of the effect $\mu_1(x) - \mu_0(x)$ for each person and averaging.

How would you compute this estimator in practice? First you can obtain predicted values under $A = 1$ and $A = 0$ separately, for everyone (regardless of actual observed treatment), then take the difference for each person, and average across people. Note that in R, the `predict` function outputs predicted values under the *observed* treatment; in contrast here one needs predicted values under $A = 1$ and $A = 0$ separately, so you need the `newdata` argument. Here is some example code for a simulated dataset:

```
> cbind(x,a,y)
                x a y
 [1,] -0.44577826 0 0
 [2,] -1.20585657 0 0
 [3,]  0.04112631 1 1
 [4,]  0.63938841 0 0
 [5,] -0.78655436 0 1
 [6,] -0.38548930 0 1
 [7,] -0.47586788 1 0
 [8,]  0.71975069 1 1
 [9,] -0.01850562 1 1
...
[100,]  2.01893816 0 1
>
> mumod <- glm(y~x+a, family=binomial)
>
> mu1hat <- predict(mumod, newdata=data.frame(x,a=1) ,type="response")
> mu0hat <- predict(mumod, newdata=data.frame(x,a=0), type="response")
>
> cbind(x,a,y, mu1hat, mu0hat, mu1hat-mu0hat)
              x a y    mu1hat    mu0hat
1   -0.44577826 0 0 0.8178912 0.5709608 0.2469305
2   -1.20585657 0 0 0.7793358 0.5113598 0.2679760
3    0.04112631 1 1 0.8397107 0.6081932 0.2315174
4    0.63938841 0 0 0.8635670 0.6522366 0.2113304
5   -0.78655436 0 1 0.8012901 0.5443888 0.2569013
6   -0.38548930 0 1 0.8207133 0.5756239 0.2450893
7   -0.47586788 1 0 0.8164699 0.5686286 0.2478412
8    0.71975069 1 1 0.8665332 0.6579775 0.2085556
9   -0.01850562 1 1 0.8371566 0.6036912 0.2334654
...
100  2.01893816 0 1 0.9073274 0.7436597 0.1636677
>
> mean(mu1hat-mu0hat)
[1] 0.2303284
```

So for the above simulated dataset, the estimated average treatment effect using logistic regression is $\widehat{\psi} = 0.23$.

Note there is no particular reason to favor logistic regression for constructing the regression estimates $\widehat{\mu}_a(x)$; one might instead consider linear regression, probit regression, regression trees, kernel estimators, splines, generalized additive models, the lasso, boosting, random forests, neural networks, deep learning, etc.

3.3.1 Properties of the Plug-in Estimator

In this section we analyze the simple plug-in estimator $\widehat{\psi}$. We do so by finding answers to three standard questions: Is the estimator consistent? What is its convergence rate? What is its asymptotic limiting distribution?

Notice that if we let $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$, then $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ and $\psi = \mathbb{E}(f)$. Thus our estimator is a sample average of an estimated function, and our target estimand is an expectation of the true function f. Thus its performance will be very closely tied to the errors in estimating the function f with $\widehat{f}$.

At this point it will be useful to take a slight detour and discuss properties of estimated functions, and introduce some new notation.

Estimated functions

First, since our analysis involves differences between the estimated function $\widehat{f}$ and the truth f, i.e., errors in estimating f with $\widehat{f}$, we will need a notion of consistency for random functions $\widehat{f}$. Recall in Chapter 1 we learned that a scalar (or Euclidean) estimator $\widehat{\psi}$ is consistent if $\widehat{\psi} - \psi = o_{\mathbb{P}}(1)$, i.e., if $\widehat{\psi}$ converges to ψ in probability.

For functions we can define an appropriate (scalar) distance measure, and then consistency will be defined as in the scalar case. Some popular distance measures for functions are:

- L_1 distance: $\|\widehat{f} - f\|_1 = \int |\widehat{f}(x) - f(x)| \, d\mathbb{P}(x)$
- L_2 distance: $\|\widehat{f} - f\|_2 = \sqrt{\int \{\widehat{f}(x) - f(x)\}^2 \, d\mathbb{P}(x)}$
- L_∞ distance: $\|\widehat{f} - f\|_\infty = \sup_{x \in \mathcal{X}} |\widehat{f}(x) - f(x)|$

Note that all of these distances are themselves random variables, since they depend on the estimated $\widehat{f}$. Now we are ready to define consistency of an estimated function.

Definition 3.1. An estimated function $\widehat{f}(x)$ is consistent for a fixed target $f(x)$ in distance measure $d(\cdot, \cdot)$ if

$$d(\widehat{f}, f) = o_{\mathbb{P}}(1).$$

Similarly, $\widehat{f}$ converges at rate $r_n \to \infty$ to f in distance d if

$$d(\widehat{f}, f) = o_{\mathbb{P}}(1/r_n).$$

In addition to having a notion of consistency or convergence for estimated functions $\widehat{f}$, it will also be useful for us to have some special notation for the expected value over a random function's argument, conditioning on the randomness in the function.

Definition 3.2. For an estimated function $\widehat{f}(x)$ built from a sample $Z^n = (Z_1, ..., Z_n)$ we use the notation

$$\mathbb{P}(\widehat{f}) = \mathbb{P}\{\widehat{f}(Z)\} \equiv \int \widehat{f}(z)\ d\mathbb{P}(z) = \mathbb{E}\left\{\widehat{f}(Z) \ \Big| \ Z^n\right\}$$

to denote expectations over a new independent observation Z, conditioning on the sample Z^n.

Remark 3.1. The heuristic interpretation of $\mathbb{P}(\widehat{f})$ is as follows: you construct the function $\widehat{f}(z)$ from a sample Z^n, and then take its average over new repeated independent draws of the argument Z. It is important to note that, for a *fixed* function $f(z)$ we have

$$\mathbb{E}\{f(Z)\} = \mathbb{P}(f)$$

whereas for a random estimated function $\widehat{f}(z)$ depending on a sample Z^n, we have

$$\mathbb{E}\{\widehat{f}(Z)\} = \mathbb{E}\left[\mathbb{E}\left\{\widehat{f}(Z) \ \Big| \ Z^n\right\}\right] \neq \mathbb{P}(\widehat{f}).$$

In particular, the quantity $\mathbb{P}(\widehat{f})$ on the right-hand-side is random (through its dependence on $\widehat{f}$ and Z^n), whereas the quantities on the left-hand-side are fixed.

Back to the standardization estimator

Now we are ready to proceed with investigating the plug-in estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ of the average treatment effect: in particular the three fundamental properties of consistency, rate of convergence, and limiting distribution.

The first tells us whether the estimator is at least converging to the correct target as sample size increases (the lowest bar we would hope an estimator would clear), the second how quickly this convergence occurs (i.e., how much information in the sample the estimator makes use of), and the third whether the estimator is well-behaved enough to give us hope for constructing confidence intervals and doing inference.

At a high level, our goal is to write $\widehat{\psi} - \psi$ as a (centered) sample average, plus some noise. We know how to analyze sample averages, since for any fixed function g of the iid observations Z, we have $(\mathbb{P}_n - \mathbb{P})g(Z) = (\mathbb{P}_n - \mathbb{E})g(Z) = O_{\mathbb{P}}(1/\sqrt{n})$ and in particular

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P})g(Z) \rightsquigarrow N\Big(0, \text{var}\{g(Z)\}\Big)$$

by the central limit theorem. Therefore the problem will be reduced to analyzing whatever the noise is.

First we will introduce a foundational decomposition for $\widehat{\psi}$ (in fact, for any estimator that takes a similar form), which will be crucial for many estimators we analyze throughout the course.

Lemma 3.1. *Let $\widehat{\psi} = \mathbb{P}_n(\widehat{f}) = \frac{1}{n}\sum_i \widehat{f}(Z_i)$ be an estimator of the generic expectation $\psi = \mathbb{P}(f) = \mathbb{E}\{f(Z)\}$ based on n samples $(Z_1, ..., Z_n)$, where $\widehat{f}$ can be any estimator and $f : \mathcal{Z} \mapsto \mathbb{R}$ any function. Then we have the decomposition*

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2 \tag{3.1}$$

where

$$\begin{aligned} Z^* &= (\mathbb{P}_n - \mathbb{P})f \\ T_1 &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) \\ T_2 &= \mathbb{P}(\widehat{f} - f). \end{aligned}$$

Proof. We have

$$\begin{aligned} \widehat{\psi} - \psi &= \mathbb{P}_n(\widehat{f}) - \mathbb{P}(f) \\ &= (\mathbb{P}_n - \mathbb{P})\widehat{f} + \mathbb{P}(\widehat{f} - f) \\ &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) + (\mathbb{P}_n - \mathbb{P})f + \mathbb{P}(\widehat{f} - f) \\ &\equiv T_1 + Z^* + T_2 \end{aligned}$$

where the first line follows by definition, the second by adding and subtracting $\mathbb{P}(\widehat{f})$ (which we recall is not the same as $\mathbb{E}(\widehat{f})$), and the third by adding and subtracting the quantity $(\mathbb{P}_n - \mathbb{P})f = (\mathbb{P}_n - \mathbb{E})f = \frac{1}{n}\sum_i [f(X_i) - \mathbb{E}\{f(X_i)\}]$. □

Remark 3.2. Lemma 3.1 immediately applies to the plug-in estimator $\widehat{\psi}$ if as before we let $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$ so that the estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ is a sample average of the estimated function $\widehat{f}$, and the target parameter $\psi = \mathbb{E}(f)$ is the population expectation of the true function f.

Lemma 3.1 has achieved our goal of writing our estimator as a centered sample average plus noise. The first term Z^* in (3.1) is a nice centered sample average, and so by the central limit theorem it behaves as a normally distributed variable with variance $\text{var}(f)/n$, up to error $o_{\mathbb{P}}(1/\sqrt{n})$. Thus our problem is reduced to analyzing the two noise terms, denoted T_1 and T_2.

Remark 3.3. Note that the decomposition in (3.1) only relied on the estimator being a sample average of an estimated function $\widehat{f}$, and on the estimand being an expectation of a true function f. There was nothing special about ψ being the average treatment effect, or f being a regression function. We will see the decomposition (3.1) repeatedly throughout the course, since many estimands are expected values of (sometimes complicated) generic functions, which can be estimated by corresponding sample averages of estimates of these functions.

Now we will analyze the noise terms T_1 and T_2.

It turns out that the term T_1 is typically of smaller order than even the Z^* term. In fact, $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ under some weak regularity conditions, as long as $\widehat{f}$ is consistent for f in L_2 norm, i.e., as long as

$$\|\widehat{f} - f\|_2^2 = \int \{\widehat{f}(x) - f(x)\}^2 \, d\mathbb{P}(x) = o_{\mathbb{P}}(1).$$

We will prove this rigorously later in the course (for a sneak peek see Kennedy et al. [2019a]). Intuitively, however, this should not be too surprising, since T_1 is a centered sample average (just like Z^*), but in fact the quantity it is averaging is shrinking to zero with n (as long as $\widehat{f}$ is tending to f). This is like taking larger and larger centered sample averages of a random variable whose variance shrinks with n.

Now we turn to the last noise term T_2, which is the really interesting one. For many estimators we discuss in the course, the T_2 term will be particularly crucial, driving the rate of convergence and limiting distribution.

3.3.2 The Parametric Plug-in Estimator

First we consider analyzing $T_2 = \mathbb{P}(\widehat{f} - f)$ in the case where $\widehat{f}$ is estimated with a (correct) parametric model, i.e., where

$$\widehat{f}(x) = f(x; \widehat{\beta})$$

for some finite-dimensional parameter $\beta \in \mathbb{R}^p$. For example, when using the logistic regression model as before, we would have $f(x;\beta) = \text{expit}(\beta_0 + \beta_1 + \beta_2^{\mathrm{T}} x) - \text{expit}(\beta_0 + \beta_2^{\mathrm{T}} x)$.

Note in the parametric case we can view

$$\begin{aligned} T_2 &= \mathbb{P}(\widehat{f} - f) \\ &= \int \{f(x;\widehat{\beta}) - f(x;\beta)\} \, d\mathbb{P}(x) \\ &\equiv g(\widehat{\beta}) - g(\beta) \end{aligned}$$

as a simple difference in functions $\widehat{\beta}$ and β, where the function g will be smooth if f is. Therefore we will first understand the error between $\widehat{\beta}$ and β, and then use the delta method.

For most smooth parametric models, the estimator $\widehat{\beta}$ will solve an estimating equation based on some mean-zero estimating function m that is smooth in β. For example, the logistic regression estimator solves an estimating equation based on the estimating function (or score function)

$$m(Z;\beta) = \begin{pmatrix} 1 \\ A \\ X \end{pmatrix} \Big\{ Y - \text{expit}(\beta_0 + \beta_1 A + \beta_2^{\mathrm{T}} X) \Big\}.$$

so that

$$\mathbb{P}_n\{m(Z;\widehat{\beta})\} = 0$$

by definition. The next standard result shows that such solutions to finite-dimensional estimating equations behave like sample averages.

Lemma 3.2. *Suppose the estimator $\widehat{\beta} \in \mathbb{R}^p$ solves an estimating equation so that*

$$\mathbb{P}_n\{m(Z;\widehat{\beta})\} = 0.$$

Assume $m(z;\beta) \in \mathbb{R}^p$ is Lipschitz in β, and that $\mathbb{E}\{m(z;\beta)\}$ is differentiable at the population β satisfying $\mathbb{E}\{m(Z;\beta)\} = 0$ with nonsingular derivative matrix. Then

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P})\left[\left\{\frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}}\right\}^{-1} m(Z;\beta)\right] + o_{\mathbb{P}}(1/\sqrt{n}), \tag{3.2}$$

Proof. See Theorem 5.23 of van der Vaart [2000]. □

Corollary 3.1. *Under the conditions of Lemma 3.2, the estimating equation estimator $\widehat{\beta}$ is root-n consistent and asymptotically normal.*

Now we have all the tools we need to analyze the quantity $\mathbb{P}(\widehat{f} - f)$ and thus the estimator $\widehat{\psi}$ in the parametric case.

Theorem 3.1. *Let $f(x) = \mu_1(x) - \mu_0(x)$ and $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$, so that $\psi = \mathbb{E}\{f(X)\}$ is the average treatement effect. Assume the parametric model*

$$\mu_a(x) = \mu_a(x;\beta)$$

for some $\beta \in \mathbb{R}^p$, and that the estimator $\widehat{\beta}$ satisfies the conditions of Lemma 3.2. Then

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z;\beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z;\beta) = f(x;\beta) + \frac{\partial \mathbb{E}\{f(X;\beta)\}}{\partial \beta^{\mathrm{T}}}\left\{\frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}}\right\}^{-1} m(z;\beta)$$

and so is root-n consistent and asymptotically normal.

Proof. By Lemma 3.1 we have

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2$$

where $Z^* = (\mathbb{P}_n - \mathbb{P})f$ and T_1 and T_2 defined accordingly. By Lemma 3.2 we have

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P})\left[\left\{\frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}}\right\}^{-1} m(Z;\beta)\right] + o_{\mathbb{P}}(1/\sqrt{n})$$

which also is enough to imply $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. Further by the delta method we have

$$\begin{aligned} T_2 &= g(\widehat{\beta}) - g(\beta) \\ &= (\mathbb{P}_n - \mathbb{P}) \left[\frac{\partial g(\beta)}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}) \end{aligned}$$

for $g(\beta) = \mathbb{E}\{f(Z;\beta)\}$. Combining the terms gives the result. □

To summarize, when $\widehat{\mu}$ is estimated with a correct parametric model, the resulting plug-in estimator $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ for $f = \mu_1 - \mu_0$ is root-n consistent for the causal effect ψ and asymptotically normal.

When the parametric model for μ is correct, this plug-in estimator is most efficient (this follows from classical low-dimensional parametric maximum likelihood theory); the intuition is that with a correct simple model we can "predict" the treatment effect much more precisely than say with the difference-in-means estimator. Further confidence intervals can be constructed using estimates of the closed-form asymptotic variance given above, or via bootstrap (which is typically easier).

Of course, these days we rarely believe our parametric models are actually correct, especially when X contains some continuous covariates or is high-dimensional. At best such models may be a modestly biased approximation, but at worst, when very misspecified they can turn our estimation procedure into garbage, yielding estimates that are not only far away from the truth, but to an unknown extent.

3.3.3 The Nonparametric Plug-in

This begs the question of how the plug-in estimator would behave if we used a more flexible estimator to construct $\widehat{\mu}$, say random forests or the lasso or deep learning.

In this case, of course the central limit theorem term Z^* in our decomposition (3.1) is still going to behave as a mean-zero normally distributed random variable with variance $\text{var}(f)/n$, since it does not depend on the estimated $\widehat{f}$. Further, even when μ is treated as a potentially infinite-dimensional function and estimated flexibly and data-adaptively, the term T_1 can still be of smaller order (though we may need to use sample splitting, as will be discussed in detail later in the course).

Unfortunately the picture is nowhere near as rosy for the important T_2 term in (3.1). If all we know about the flexible estimator $\widehat{f}$ are high-level rates of convergence, say in L_2 norm, then all we can say about T_2 is

$$T_2 = \mathbb{P}(\widehat{f} - f) \leq \sqrt{\mathbb{P}\{(\widehat{f} - f)^2\}} = \|\widehat{f} - f\|_2$$

where the second inequality uses Cauchy-Schwarz. This means in general we would expect the plug-in estimator $\widehat{\psi}$ to *inherit* the (typically slow) rate of convergence of the nonparametric estimator $\widehat{f}$.

This is a problem since for most realistic infinite-dimensional function classes the L_2 norm will be far away from $1/\sqrt{n}$. For example when f lies in a Hölder class with index s (i.e., all partial derivatives up to order s exist and s^{th} derivatives Lipschitz) then for *any* estimator $\widehat{f}$ the rate cannot be any faster than

$$\|\widehat{f} - f\|_2 \ \gtrsim \ n^{-s/(2s+d)}$$

uniformly over the Hölder class [Tsybakov, 2009]; note this rate is always slower than $\sqrt{n}$. For example, suppose we are only willing to assume our regression functions have $s = 2$ derivatives, and we have $d = 16$ covariates. Then the best achievable rate is $n^{-1/10}$. Neural network classes are known for yielding dimension-independent rates [Györfi et al., 2002], but even these are $n^{-1/4}$, a far cry from $1/\sqrt{n}$.

Further, when $\widehat{\mu}$ is estimated flexibly with modern nonparametric tools, we do not only pay a price in the rate of convergence – it will generally also be true that, even if we can derive a tractable limiting distribution, there will be some smoothing bias, so confidence intervals will not be correctly centered (even using the bootstrap) and thus will not cover at the nominal level. However, often complex nonparametric estimators do not even yield tractable limiting distributions, even uncentered.

Here we find ourselves in a bit of a quandary. We could use the simple difference-in-means estimator, which is root-n consistent and asymptotically normal *under no modeling assumptions*; however it completely ignores covariate information and so will be very inefficient relative to other estimators. Alternatively we could model the regression function and use the plug-in estimator. However if we use parametric models to achieve root-n rates and small confidence intervals, we are putting ourselves at great risk of bias due to model misspecification; on the other hand, if we model the regression functions nonparametrically, letting the data speak for themselves, then we will typically suffer from the curse of dimensionality and be subject to slow rates of convergence, and at a loss for confidence intervals and inference.

What should we do? Is there any way to get the best of both worlds, using the covariates to gain efficiency over the difference-in-means estimator, but retaining its model-free benefits and not risking bias?

Bibliography

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.